A French Anonymization Experiment with Health Data

Maxime Bergeat, Nora Cuppens-Bouhlahia, Frédéric Cuppens, Noémie Jess, Françoise Dupont, Said Oulmakhzoune, Gaël de Peretti

A French Anonymization Experiment with Health Data

Maxime Bergeat[*], Nora Cuppens-Boulahia[**], Frédéric Cuppens[**], Noémie Jess[***], Françoise Dupont[****], Saïd Oulmakhzoune[**], Gaël de Peretti[*]

* French National Institute for Statistics and Economic Studies (Insee), Paris, maxime.bergeat@insee.fr, gael.de-peretti@insee.fr
** Institut Mines Télécom - Télécom Bretagne, Rennes, nora.cuppens@telecom-bretagne.eu, frederic.cuppens@telecom-bretagne.eu, said.oulmakhzoune@telecom-bretagne.eu
*** French Ministry of Health and Solidarity, Statistical Department, Paris, noemie.jess@sante.gouv.fr
**** French Secure Remote Data Access Centre (CASD), Paris, francoise.dupont@casd.eu

Abstract: In this paper, a case study about a microdata anonymization test is presented. The work has been made considering a French administrative health dataset with indirect identifiers and sensitive variables about hospital stays. Two approaches to build a *k*-anonymized file are described, and software used in the test are compared.
Keywords: case study, health data, *k*-anonymization, microdata, statistical disclosure control

1 Introduction

Following the "Free Health Data" petition launched by the French Inter-Associative Health Group CISS (*Collectif Interassociatif Sur la Santé*) in January 2013, the French Minister of Health and Solidarity Marisol Touraine asked P.L. Bras to write a report on health data use and governance. It was released in September 2013. The debate on increasing access to public health data was launched in late November 2013. M. Touraine commissioned F. Von Lennep, head of the statistical department of the Ministry of Health and Solidarity, and P. Burnel, responsible for IT System Strategies in the Ministry of Health and Solidarity, to lead the Open Data committee. The committee issued its conclusions in late April 2014.

A taskforce leaded by André Loth, project manager in the statistical department of the Ministry of Health and Solidarity, was launched to assess data safety and individual re-identification risk in health data. Its work focused on health data specifics, anonymization techniques, and how a wider access to health data could be organized. In order to get practical perspectives, the taskforce launched a test on real data. The aim was to determine how anonymized data files could be built from the original French exhaustive administrative database on hospital stays. The objective of this work is to specify a suitable method for producing anonymized microdata files to be released as public use files. A wide range of different users is expected to use these data files: from citizens to health insurers, including pharmaceutical companies, doctors…

Our work fell into a binding timetable due to organizational and legal constraints. The French Commission for Data Protection and Liberties (*Commission Nationale de l'Informatique et des Libertés*) gave its agreement on January 30th, 2014, for a three-month renewable period. The test began in early March 2014 because of late availability of data and needed nomenclatures. Test conclusions were given in late April 2014.

This paper aims at presenting and comparing two ways to elaborate a safe microdata file on this case study. This paper is organized as follows. Section 2 gives an overview of the software used for the test, µ-Argus and ARX. In Section 3 test data are introduced and we present objectives of the anonymization process in terms of disclosure risk and allowed technical solutions to reduce it. Section 4 presents two approaches tested to reach anonymization goals. A discussion is led in Section 5 and some conclusions are given in Section 6.

2 Data Anonymization Tools

The result of de-identification tools market browsing has shown that there is a wide range of solutions developed for structured data anonymization offering a long list of functionalities. But few of them are used or can be smartly used (because they are internal to specific organizations or developed by researchers for specific use).

Moreover, practical applicability of the promising functions needs to be tested, in particular to highlight what has to be refined, updated or further developed to enhance the usability.

Of course these tools are different from those used for masking data that we do not consider in this paper, because they only manipulate direct identifiers by creating pseudonyms or applying randomization techniques, but never deal with indirect identifiers and thus do not provide suitable protection.
Some of those de-identification tools and some of their features are summarized in Table A.1 in Appendix.

We can observe that there are few interesting commercial products whereas there are several tools and prototypes developed by academic researchers not broadly used. To further study the usability, the robustness of these tools and privacy protection they can provide, a testing process on real data of significant size needs to be performed. In the two next subsections, we will introduce the software used in the test, µ-Argus and ARX. Further comparison is also given in Section 5.

2.1 µ-Argus

µ-Argus is a software package developed to help statisticians to anonymize microdata files. It is a deliverable of the European CASC (Computational Aspects of Statistical Confidentiality) project that took place between 2000 and 2003, and additional work has been made since 2003 thanks to other European projects: the Centre of Excellence SDC (CENEX-SDC), ESSNet-harmonization and ESSNet-SDC. The next major release of µ-Argus will be an open-source version; development is still in progress (see de Wolf (2013) for more details).

Many protection methods can be used with µ-Argus. The anonymization scheme can be summarized in a few steps:

- Import microdata into µ-Argus and define metadata before beginning the statistical disclosure control (SDC) process. In this step we must for instance define the indirect identifiers and the sensitive variables of the microdata file.
- Estimation of the disclosure risk of the file. It is possible to set a *minimum* threshold for specified combinations of values of indirect identifiers. In the risk estimation you can also, in case of a survey, use the sampling weights in order to probabilistically estimate the re-identification risk.
- Several protection methods to reduce disclosure risk are implemented in the software: they are perturbative (noise addition, microaggregation, post-randomization method (PRAM), rounding, data swapping…) or not (global recoding, local suppression…). For a further description of µ-Argus features see Hundepool (2008).
- After residual disclosure risk estimation, safe microdata files can be exported.

µ-Argus is an interactive program mostly used by National Statistical Institutes (NSIs). A questionnaire on SDC tools led in the European NSIs in 2013 has shown that µ-Argus is often used for microdata protection in an automated way, even if some institutes are still applying manual procedures or homemade tools.

2.2 ARX

ARX Data Anonymization Tool is an open source data anonymization framework developed in Java. It implements dedicated optimizations for anonymization algorithms. It encodes data in a way that allows evaluating transformations efficiently. ARX is available online[1]. See also Kohlmayer and Prasser (2012).

ARX is a flexible and intuitive high-performance data anonymization tool. It implements a wide variety of common privacy criteria such as *k*-anonymity, *l*-diversity, *t*-closeness, δ-presence and their combinations. See Sweeney (2002), Machanavajjhala (2007), Li (2007) and Nergiz (2007) for more information about these concepts. ARX allows protecting a dataset from a multitude of privacy threats. It allows non-domain-experts to specify privacy guarantees in an intuitive manner and to explore potentially large solution spaces efficiently. It implements also a set of metrics to evaluate information loss.

The ARX graphical tool is composed of three main perspectives (views). The first one allows defining the required privacy guarantees. The second perspective allows browsing the complete solution space according to the privacy criteria defined on the first perspective. Solutions that fulfill the criteria as well as the "optimal" solution are highlighted. The third perspective allows comparing transformed datasets to the original dataset.

[1] It can be downloaded from http://arx.deidentifier.org

The ARX anonymization scheme could be summarized in the following steps:

- Import the original dataset in CSV format.
- Define direct identifiers, indirect identifiers and sensitive variables of the original dataset.
- Define the privacy criteria. For instance 10-anonymity for indirect identifiers, 3-diversity for the sensitive variable, …
- Define generalization hierarchies for each indirect identifier. ARX allows importing generalization hierarchies (nomenclatures) or creating new ones from an interactive menu.
- Define the maximum number of outliers that can be tolerated and will be suppressed.
- Start the anonymization process.
- Browse and analyze solutions.
- Export desired transformations.

3 The Test Framework

3.1 The Dataset to Anonymize

The anonymization test is performed with the French dataset PMSI (*Programme de Médicalisation des Systèmes d'Information*). This exhaustive medical-administrative file contains more than 35 million records corresponding to all hospital stays for one calendar year in the three major hospitalization areas: MCO (*médecine, chirurgie, obstétrique* or *MSO*: medicine, surgery and obstetrics) also known as short stays, rehabilitation care and home hospitalization.

For our test, consistently with the desired output file, we experiment on the latest PMSI-MCO file that contains short stays. Besides, the working group decides to focus on "unique" stays, in other words hospital stays except medical sessions (*e.g.* chemotherapy or dialysis sessions) that generally happen more than once a year. The underlying reason is that there are different ways for encoding doctors to fill in information about medical sessions: they are allowed to create one record for each stay or one for all. As such, interpretation of these stays is hard and it is difficult to deal with them in the anonymization process: for instance the variable "length of stay" is very volatile because of the different possibilities for health workers to fill in data: this can create fake anonymization issues.

All things considered, we deal with the 2012 PMSI-MCO file excluding medical sessions. The dataset contains 20.6 million hospital stays. For each stay detailed medical information (notably major diagnosis, related and accompanying diagnoses, medical procedures, care areas) and administrative information (hospital identification number FINESS[2], dates of stay, admission and discharge modes, patient birth date, sex and residential area) are provided. The FINESS number is public information since it is released on a dedicated website[3], and

[2] *FIchier National des Établissements Sanitaires et Sociaux* or National file of the sanitory and social establishments

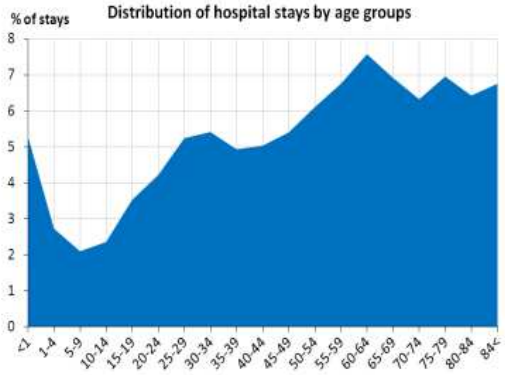
[3] See http://finess.sante.gouv.fr/jsp/index.jsp

enables one to find the name, location, legal status and type of a hospital. Each record (*i.e.* each stay) is classified into one GHM[4] code (the French version of DRG, Diagnosis Related Group) that is derived from medical data and in a way sums up the pathology for which the patient is hospitalized. The most aggregated level of the GHM classification is the CMD[5] (Major category of diagnosis), which is a 28-item nomenclature including one code for medical sessions and one error code.

First of all we asked health experts what the most used classification for health studies currently is, in order to maximize data utility of anonymized files and to ensure consistency with other datasets already disseminated. For example, it is very common to classify age in five-year age groups.

Before beginning the anonymization process, we have analyzed our dataset (distributions of the variables, health experts requirements) in order to see the variables (and their related items) that are rare and as such likely to lie behind anonymization issues. Hospital stays are uncommon among children between 1 and 14 years old in particular (see Figure 1 below). Concerning admission and discharge variables, origins and destinations other than home, unknown or regular transfers are quite unusual and death is in addition particularly sensitive (see Figure 1 below). Likewise, we find a few long hospital stays (more than one week). Lastly, distance between residential and hospitalization areas has a strong power of re-identification: patients whose residential area is not in the same territory as their hospital are in general easy to disclose. We take one random example with the first (by alphabetic order) French department (*Ain*) with population of about 600 000 inhabitants: for 43% of hospital stays in *Ain* the patient also lives in the department. Conversely, 86 combinations of *Ain-other department* (out of a total of 100 combinations) consist in less than 10 hospital stays.

[4] *Groupe Homogène de Maladie*
[5] *Catégorie Majeure de Diagnostic*



Patients aged:
→1-4: 2.7% of hospital stays
→5-9: 2.1% of hospital stays
→10-14: 2.4% of hospital stays

Vs. around 6% of hospital stays for each other age group

Distribution of hospital stays by mode of		
	Admission	Discharge
Home	83.1%	77.3%
Unknown	14.5%	14.5%
Regular transfer	2.1%	5.4%
Change	**0.2%**	**1.3%**
Other transfert	**0.1%**	**0.1%**
Death	-	**1.4%**

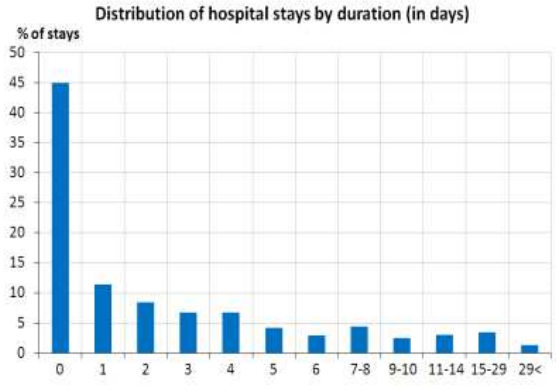


Figure 1 Univariate analyses about indirect identifiers

3.2 Objectives of the Anonymization Process

The anonymization task consists in obtaining a dataset that reaches *k*-anonymity and *l*-diversity given that we address exhaustive data. We recall the definitions of these concepts below.

Definition 1 *k*-anonymity
A file is said to be *k*-anonymized if at least *k* records match every combination of values taken by the indirect identifiers. Each combination is called an identification key.

Definition 2 *l*-diversity
A file is said to be *l*-diverse if for each identification key, there are at least *l* "well represented" different values for each sensitive attribute.

The taskforce about the anonymization test decided that neither perturbative methods nor local suppression were suitable in order to obtain the safe file. We have considered that using perturbative methods may increase risk of misuse, especially with non-specialist users. Moreover we agreed on the idea that local suppression may lead to bad interpretation of the data so local suppression was not allowed either. In the ARX scenario, outliers with a high

risk of re-identification are computed, and there is post-processing in order to deal with them without deleting these risky records or using local suppression: see also Section 4.2.

We use the following variables as indirect identifiers:

- Sex
- Age
- Residential area (ZIP code)
- Hospitalization area (FINESS number)
- Length of hospital stay

For the µ-Argus scenario, we consider two other indirect identifiers to ensure *k*-anonymity:

- Admission mode
- Discharge mode

We also check the variable "Major category of diagnosis" (CMD: *Catégorie Majeure de Diagnostic*) for *l*-diversity. This is an aggregated version of the main diagnosis recorded during the patient stay. This variable has 26 modalities for all stays considered in the anonymization test. As previously indicated, repeated medical sessions are excluded from the test, there is a specific CMD code for medical sessions that we do not consider afterwards. There is no error code in the dataset.

The objective of the anonymization process is to obtain at the end a 10-anonymised and 3-diverse file. A modality is said "well represented" (for *l*-diversity) if there is at least one record with the considered modality. Definition of such levels was done after a long discussion: these levels were decided in a way both cautious and reasonable… A value of k= 5 is often used for some health data dissemination; we chose k= 10 to ensure a security margin, for instance if the potential attacker uses auxiliary information to deconstruct the anonymization process. Moreover, it is theoretically easier to get 3-diversity with a 10-anonymized file than a 5-anonymized one. For more details about the choice of *k* in the SDC process see El Emam (2009). The choice of l= 3 aims at avoiding exact disclosure of the sensitive variable "Major category of diagnosis".

4 Results of the Experiment

Two ways to create a safe file were proposed during the test. In this section we will present two different approaches to elaborate a file regarded as safe given the anonymization goals introduced in Section 3. We will in the next Section compare the two methods and also software used to build the anonymized files.

Given we want to obtain a safe file without any local suppression or use of perturbative methods, the only approach to reach *k*-anonymization and *l*-diversity is global recoding. The first step in preliminary work was to discuss with users of the PMSI file and domain experts in order to have predefined ways of recoding. The goal was to have an overview of "smart" plausible recoding if an expert in health data wants to make statistical studies with the protected file. For instance, studying children under 1 year old is a classical subject in pediatrics. The definition of possible ways of recoding is completed with the analysis of distributions of the indirect identifiers.

We also have reached the conclusion in the preliminary work that sampling should give additional protection in order to avoid residual disclosure risk. Yet the completeness of hospital stays is a major asset of the French PMSI file.

4.1 µ-Argus: Using Global Recoding with an Iterative Approach

In the first test scenario we produced safe files with µ-Argus. Here are the different steps:

- After discussion with domain experts, definition of multiple predefined ways of recoding.
- Calculation of frequencies for all identification keys. In order to limit the number of identification keys to compute some variables have to be discretized (*e.g.* age has been recoded to larger bands with a top coding, and ZIP codes are replaced by department - NUTS3 level).
- For each modality of each variable, µ-Argus computes the number of identification keys with insufficient frequency where this modality is implied. According to these results we use a step-by-step approach:
- While the file is not 10-anonymized, do:
    - Detect modalities that are implied in a big number of identification keys with less than 10 stays. Use the ratio between number of concerned identification keys and frequency of occurrence (the two numbers are given by µ-Argus) to detect "risky" modalities.
    - If it is possible from an analytic point of view given conclusions of discussion with health experts, combine this "risky" modality with an other one.
    - If not possible (for instance there is no sense in combining two geographical places that are not in the same area!), make another recoding: go back to the detection step.
- When you obtain a 10-anonymized file, check for 3-diversity using a SAS algorithm.

We have constructed two files using this empirical approach. Two files that are compatible with anonymization goals are described in Table 1 and Table 2. In the file of Table 1 there is no geographical dimension: we have tried to introduce the variable "Residential area" and we have obtained the file summarized in Table 2.

Name of variable	**Type of variable**	**Number of modalities**
Sex	Indirect identifier	2
Age	Indirect identifier	18
Length of hospital stay	Indirect identifier	12
Admission mode	Indirect identifier	2
Discharge mode	Indirect identifier	2
Major category of diagnosis	Sensitive variable	26

Table 1 One file that meets anonymization criteria without geographical dimension

Name of variable	Type of variable	Number of modalities
Sex	Indirect identifier	2
Age	Indirect identifier	6
Residential area	Indirect identifier	22 (NUTS2 level grouping overseas departments and combining Corsica and PACA[6] regions)
Length of hospital stay	Indirect identifier	2
Admission mode	Indirect identifier	2
Discharge mode	Indirect identifier	2
Major category of diagnosis	Sensitive variable	26

Table 2 One file that meets anonymization criteria with one geographical dimension

The variable "hospitalization area" is not included in the two proposed files: it is hard to build such *l*-diverse file because some hospitals are specialized in the treatment of a particular disease, for instance cancer centers. Moreover a file with two geographical levels will *a priori* be very risky because hospital stays where residence and hospitalization areas are different are extremely rare: like already described in Section 3, this can easily lead to re-identification. However we didn't try to consider instead "hospitalization area" a variable that takes the modalities "hospitalization area is close to residential area" and "hospitalization area is far away from residential area". This binary variable should reach a quite good trade-off between data utility and disclosure risk; it has not been tested here because of hard work needed to compute this variable.

4.2 ARX: Two Levels of Detail Depending on Disclosure Risk of the Stay

The second test has been implemented with the ARX Data Anonymization Tool. Privacy criteria are also 10-anonymity and 3-diversity. In ARX, we define first possible generalizations (*i.e.* ways of recoding) for each indirect identifier. Then ARX chooses a possible combination of those generalizations that fulfills the given privacy criteria.

We use the following five variables as indirect identifiers:

- Sex
- Age
- ZIP code of residential area
- FINESS number
- Length of hospital stay

Note that admission and discharge modes are not studied in this anonymization scenario.

The different steps of the test scenario could be summarized as follows:

- Creation and loading of the dataset input. The ARX input file should be in CSV format.

[6] *Provence-Alpes-Côte d'Azur* is a French region located in the south and close to Corsica.

- After loading the original dataset, indirect identifiers and sensitive attributes are defined. Indirect identifiers will be *k*-anonymized, and sensitive attributes will be *l*-diversified.
- For each indirect identifier attribute we import its corresponding generalization hierarchy, which defines a way of iteratively generalizing the values of an attribute. The generalization hierarchy should be monotonic, *i.e.* within one hierarchy the groups at level *n+1* are built by merging groups from level *n* (see for instance Table A.2 and Table A.3 in Appendix). All levels of generalization are represented in Table 3.
- Definition of anonymization criteria:
    - 10-anonymity for the five indirect identifiers
    - 3-diversity for the sensitive attribute "Major category of diagnosis"
    - Suppression threshold (percentage of outliers that can be globally suppressed)
- Production of anonymized solutions. ARX searches all possible solutions by applying systematically all possible combinations of generalization hierarchies of indirect identifiers to check for anonymization criteria.

Indirect identifier	**Level 0**	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Sex	Cleartext	Not disseminated	Not disseminated	Not disseminated	Not disseminated
Age (years)	Cleartext	Five-year age groups (see Figure A.2)	Level 2 (see Figure A.2)	Level 3 (see Figure A.2)	Not disseminated
ZIP code of residence	Cleartext	NUTS3 (101 departments in France)	NUTS2 (22 regions in Metropolitan France and overseas departments)	NUTS2 level grouping overseas departments and combining Corsica and PACA regions	Not disseminated
FINESS number (hospitalization area)	Cleartext	NUTS3 (101 departments in France)	NUTS2 (22 regions in Metropolitan France and overseas departments)	NUTS2 level grouping overseas departments and combining Corsica and PACA regions	Not disseminated
Length of hospital stay (days)	Cleartext	Level 1 (see Figure A.3)	Level 2 (see Figure A.3)	Not disseminated	Not disseminated

Table 3 Generalization levels in the ARX scenario

ARX allows defining tolerated suppression threshold *i.e.* the percentage of outliers (records allowed to be suppressed in order to fulfill the anonymization criteria) that are (globally) suppressed.

Firstly we calculate possible solutions for the given anonymization criteria with no suppression (threshold= 0%). We obtain 127 solutions among 1000 combinations of generalization hierarchies (2 x 5 x 5 x 5 x 4). There is no solution where at least one indirect identifier is not completely anonymized (the variable is suppressed). Like for the first test scenario, solutions with no suppression suffer from a high loss of utility. Table 4 gives an overview of a possible solution. Information about "hospitalization area" and "length of hospital stay" are not disseminated in order to fulfill anonymization criteria.

Name of variable	**Type of variable**	**Level of generalization**
Sex	Indirect identifier	Cleartext
Age	Indirect identifier	Five-year age groups (level 1)
Residential area	Indirect identifier	NUTS3 level
Major category of diagnosis	Sensitive variable	Cleartext

Table 4 One combination that meets anonymization criteria in the ARX scenario

It is therefore interesting to create files with consideration of the outliers. To do this, we apply the following methodology in order to obtain a 10-anonymized and 3-diverse file:

- Set the *maximum* threshold of (globally suppressed) outliers to a positive value (4% in this test), and then select one solution.
- Extract outliers, *i.e.* a set of records that do not fulfill the anonymization criteria
- Reprocessing deleted records. It corresponds to replay the anonymization process on deleted records.
- Construct the final solution. We can integrate deleted records after anonymization, and then we obtain a new file with two levels of detail depending of the rareness of each stay.

In this test, *maximum* threshold of outliers is fixed to 4%.

Table 5 illustrates an example of solution corresponding to the threshold 3.8% (650 388 records are outliers). The first step is to extract outliers corresponding to one solution from the original dataset. Then we anonymize outliers by applying the same process. It is obvious that information provided for outliers is less detailed. This approach differs slightly from a method with local suppression because here the information provided is the same for all outliers.

Name of variable	Type of variable	Level of generalization	
		For the outliers	**For other stays**
Sex	Indirect identifier	Cleartext	Cleartext
Age	Indirect identifier	Five-year age groups (level 1)	Five-year age groups (level 1)
Residential area	Indirect identifier	NUTS3 level	NUTS3 level
FINESS number	Indirect identifier	Not disseminated	NUTS2 level
Length of hospital stay	Indirect identifier	Not disseminated	Level 1
Major category of diagnosis	Sensitive variable	Cleartext	Cleartext

Table 5 One file that meets anonymization goals with two levels of detail depending on the rareness of a stay

The file described in Table 5 has strictly more utility than the previous one because information provided for non-outliers is more detailed: data about place of hospitalization and duration of the stay may be disseminated for records with a low risk of re-identification.

5 Discussion

Given the non-uniform distribution of indirect identifiers shown in Section 3, building *k*-anonymized files is very hard without using any perturbative method because of the rareness of some modalities of indirect identifiers. See Sweeney (2000) for another example with demographic data. The loss of utility is very high when making global recoding. In particular, we didn't manage to obtain safe files with two geographical dimensions (residence and hospitalization areas). Even in the file described in Table 5, there is no information about place of hospitalization for the outliers.

Both software of this test are quite easy to use with an intuitive interface and handle very large datasets like the French PMSI file - more than 20 million records. After generation of the microdata file and computation of frequencies for all identification keys that are time-consuming, other steps of the anonymization process are done instantly with µ-Argus. The anonymization process (comparison of all possible combinations of generalization hierarchies) consists in less than 2 minutes with ARX.

µ-Argus is a flexible program where all parameters of the SDC process can be controlled, there is no "black-box" effect and the documentation about the software is detailed. ARX has apparently more functionalities and anonymization criteria than µ-Argus like *l*-diversity (that can be checked directly during the SDC process), *t*-closeness and δ-presence. However documentation is limited and it is sometimes hard to use some functionalities: in particular, there is only little documentation about metrics used to choose one generalization hierarchy by minimizing information loss.

Some steps are not optimized in µ-Argus and ARX. For instance there is no way to compute optimal global recoding in order to minimize information loss under the constraint of getting a *k*-anonymized file. Some algorithms are described in the literature (see Lefevre (2006) about the Mondrian algorithm) but they are not implemented in the software used in this test.

Moreover it is not easy to add utility constraints when using this kind of algorithm. For instance, we do not want to recode a geographical variable by regrouping some areas that are very far away from each other. That's why all methods presented in this paper consist in iterative empirical approaches in order to reach *k*-anonymization and *l*-diversity.

6 Concluding remarks

We can see in this study that it is hard to reach a good trade-off between disclosure risk and data utility when protecting a microdata file using neither perturbative methods nor local suppression. All files proposed in this test for dissemination suffer from a huge loss of data utility, despite the large number of records in the original PMSI file of hospital stays in 2012. Two methods to obtain a 10-anonymized and 3-diverse file were presented: the former with "pure" global recoding, the latter with different levels of generalization according to the power of re-identification of each record.

In further work it should be interesting to analyze disclosure risk and associated loss of information when perturbing data, for instance with qualitative microaggregation (see also Hundepool (2012) for a definition of microaggregation). Using geographical information for perturbation seems to be a good idea because it is easier with this kind of variable to control the perturbation and the associated loss of information compared to a non-ordinal variable like patient sex. Lastly, values of *k* and *l* were fixed in this study: it might be of interest to perform tests with other values and study resulting data utility and residual disclosure risk.

References

Bras, P.L. & Loth, A. (2013). *Rapport sur la gouvernance et l'utilisation des données de santé*. Available online[7].

El Emam, K. & *al*. (2009). *A Globally Optimal k-Anonymity Method for the De-Identification of Health Data*, Journal of the American Medical Informatics Association.

Eurostat (2013). *Results of the questionnaire on SDC tools*, proceedings of the fifth meeting of the Expert Group on Statistical Disclosure Control.

Hundepool, A. & *al*. (2008). *μ-Argus User's Manual.* Available online.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K. & de Wolf, P.P. (2012). *Statistical disclosure control*, Wiley Series in Survey Methodology.

Kohlmayer, F., Prasser, F. & *al*. (2012). *Highly efficient optimal k-anonymity for biomedical datasets*, CBMS.

Lefevre, K. & *al*. (2006). *Mondrian Multidimensional k-Anonymity*, ICDE.

Li, N. & al. (2007). *t-closeness: privacy beyond k-anonymity and l-diversity*, ICDE.

Machanavajjhala, A., & al. (2007). *l-diversity: privacy beyond k-anonymity*, TKDD, 1, 1, 3.

Nergiz, M. & al. (2007). *Hiding the presence of individuals from shared databases*, SIGMOD.

Sweeney, L. (2000). *Simple Demographics Often Identify People Uniquely*, Data Privacy working paper.

Sweeney, L. (2002). *k-anonymity: A model for protecting privacy*, IJUFKS, 10, 5, 557-570.

de Wolf, P.P. (2013). *Open source software Argus*, UNECE Work session on statistical data confidentiality.

[7] See http://www.drees.sante.gouv.fr/rapport-sur-la-gouvernance-et-l-utilisation-des-donnees-de,11202.html

Appendix

Tool	Issuer	Open source	Main functions	HIPAA compliance	Frequent updates	Observations
µ-Argus	Statistics Netherlands	No (a next release will be open source)	Global recoding Local suppression Top and bottom coding Post Randomization Additive noise Microaggregation Numerical Rank Swapping	No	No	Widely used by National Statistical Institutes, several perturbative and not perturbative protection methods implemented See also Section 2.1
ARX	Munich University	Yes	*k*-anonymity, *l*-diversity, *t*-closeness and *δ*-presence	No	Yes	Helps in the creation of generalization hierarchies, helps in exploring the domain of solutions, helps in comparing original data with transformed one See also Section 2.2
sdcMicro[8]	Vienna University of technology	Yes	Microaggregation, noise, swapping, local suppression, partially synthetic data	No	Yes	Cannot handle extremely large datasets, test-oriented tool, it is a R package
PARAT[9]	Privacy Analytics Inc.	No, but demo is available after requesting access and registration	Implements comprehensive risk management for three types of identity disclosure risk, evaluates data utility, suppression, data shifting across several databases,	Yes	Yes	Combines masking and de-identification techniques, Simulates attacks to determine levels of risk associated with the re-identification,

[8] Available online at http://cran.r-project.org/web/packages/sdcMicro
[9] Available online at http://www.himss.org/News/NewsDetail.aspx?ItemNumber=29971

			etc.			Integrates two expert systems (ERB & IRB) to assist risk evaluation
CAT[10]	Cornell University	Yes	*k*-anonymity, Incognito, *l*-diversity, risk analyzer	No	No	Usability problem, lack of documentation, buggy tool
UTD[11]	University of Texas (Dallas)	Yes	*k*-anonymity, Datafly, Mondrian Multidimensional, Incognito, Incognito with *l*-diversity and *t*-closeness	No	No	More developer oriented than final user oriented tool, Java implementations
And several other tools: Oracle Data Masking, Camouflage, Informatica Data Privacy, Data Masker or IBM Optim Data Privacy Solution, *etc*.						

Table A.1 Some available de-identification tools.

Level	**Age (years)**																**Sex**	
4	*																*	
3	[0-39]												...	...	...		*	
2	0	[1-4]			[5-14]						...	...	...	...	...	...	*	
1	0	[1-4]			[5-9]			[10-14]			...	...	...	...	...	...	*	
0	0	1	…	4	5	…	9	10	…	14	15	…	39	40	…	…	1	2

Table A.2 Examples of generalization hierarchy: attributes "Age" and "Sex"

Level	**Length of hospital stay (days)**																	
4	*																	
3	*																	
2	*0*	[1-2]		…	[5-6]		[7-10]				+11							
1	*0*	1	2	…	5	6	[7-8]		[9-10]		[11-14]			[15-29]			+30	
0	*0*	1	2	…	5	6	7	8	9	10	11	…	14	15	…	29	30	…

Table A.3 Example of generalization hierarchy: attribute "Length of hospital stay"

[10] Available online at http://sourceforge.net/projects/anony-toolkit
[11] Available online at http://cs.utdallas.edu/dspl/cgi-bin/toolbox